Texture and shape in fluorescence pattern identification for auto-immune disease diagnosis

V. Snell, W. Christmas, J. Kittler
CVSSP, University of Surrey, Guildford, UK
v.snell@surrey.ac.uk

Abstract

Automation of HEp-2 cell pattern classification would drastically improve the accuracy and throughput of diagnostic services for many auto-immune diseases, but it has proven difficult to reach a sufficient level of precision. Correct diagnosis relies on a subtle assessment of texture type in microscopic images of indirect immunofluorescence (IIF), which so far has eluded reliable replication through automated measurements. We introduce a combination of spectral analysis and multi-scale digital filtering to extract the most discriminative variables from the cell images. We also apply the most powerful classification techniques to make optimal use of the limited labelled data set. Overall error rate of 1.6% is achieved in recognition of 6 different cell patterns, which drops to 0.5% if only positive samples are considered.

1. Introduction

The HEp-2 Cells Classification contest at ICPR 2012 is aimed at improving the performance of auto-immune disease diagnosis through automated laboratory systems. A wide variety of these diseases affect different parts of the body, but are all associated with an immune reaction to, and an attack on, the person's own tissues. This reaction, known as Anti-nuclear antibody (ANA), can be visualised using indirect immunofluorescence (IIF), most commonly utilising the HEp-2 cell line, and forms the most reliable basis for ascertaining the presence of, and establishing the specific type of auto-immune disease. The diagnosis is usually performed by highly trained physicians directly at the microscope, although better results can be obtained through digital imaging of the slides, as the fluorescence decays fairly rapidly. Both the overall brightness and the visual pattern of the fluorescence feed into the diagnostic decision, although many clinical settings will only use the brighter samples, known as positive, for identification of specific patterns.

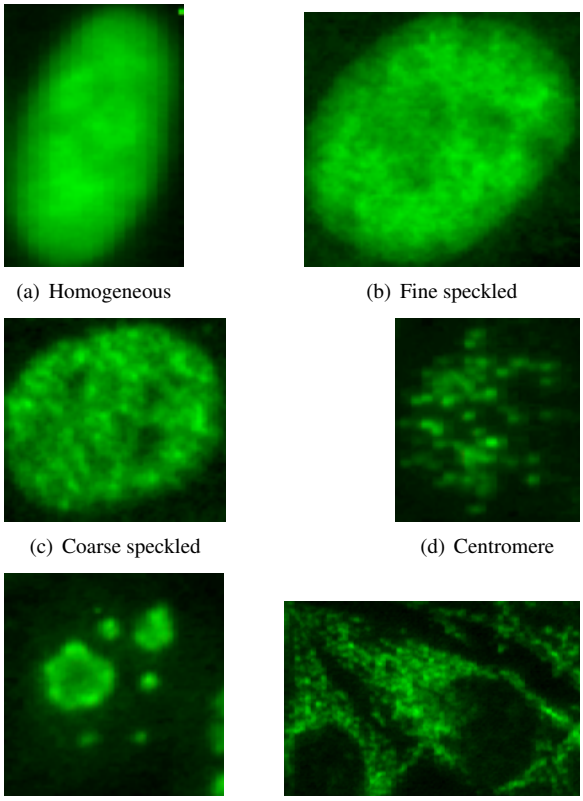
A large number of these visual patterns of fluorescence is described in the medical literature, and various groups or subsets of these have been targeted for automatic recognition by previous published works in the computer vision field. The contest provides a consistent base-line for comparing potential approaches to solving this problem by supplying a public data set, which is described below.

2. Data set

The contest training data consists of 721 images of individual cell IIF patterns, each having an associated binary mask, and intensity label (*positive* or *intermediate*), and a ground-truth class label from one of 6 classes. Therefore issues of segmentation are outside the contest scope. The classes are approximately, though not precisely, balanced, and roughly equal numbers of positive and intermediate examples of each class are included. The classes are as follows:

- **Homogeneous**: a diffuse pattern, fairly uniform across the whole nucleus.
- **Fine speckled**: a very fine-grained isotropic texture, not dissimilar to white noise.
- **Coarse speckled**: an isotropic texture of somewhat larger specks.
- **Centromere**: this class is characterised by large numbers of strong bright spots on a darker background. These are 2-3 pixels across, and 40-60 are supposed to be present, although in a number of intermediate intensity examples of this class none are visible to the eye, even after contrast normalisation.
- **Nucleolar**: a small number (less than 6) of larger bright areas within the nucleus.
- **Cytoplasmatic**: these nuclei are characterised by a strongly irregular shape, as compared to the gener-

ally elliptic nature of all other classes. The texture is equally irregular.



(a) Homogeneous (b) Fine speckled

(c) Coarse speckled (d) Centromere

(e) Nucleolar (f) Cytoplasmatic

Figure 1. Positive examples of each class

Examples of each class are given in Fig. 1, contrast boosted to make their features visible in print. Typical contrast range for positive examples is around 120 grey-levels, but can be as low as 25 for intermediate samples, greatly exacerbating the effect of sensor noise. Images sizes range from 45 to 130 pixels across.

Test images are withheld by the contest organisers, with trained classifiers submitted by contest participants. Therefore only training cross-validation results can be reported here.

3. Previous works

Computer vision researchers have attempted to automate classification of ANA IIF patterns for several years now. Although it is difficult to compare their results directly, as they use different private data sets and variable class definitions, the error rates for identification of individual cell patterns range 10-25% [5, 3, 2]. This is a promising start for early works, but not sufficient for widespread clinical application.

Study of prior publications on this topic points to two areas of potential improvement: formulation of suitable features, and use of more powerful classifiers. As should be clear from the class descriptions above, most distinctions between IIF patters are based on texture, yet texture measures form a small part of the feature sets used in prior works, and an even smaller part of the discussion. We believe that finding the appropriate texture measurements is key to solving this entire problem.

Similarly, only relatively simple classifier types, such as K-nearest neighbour and neural networks, have been brought to bear on this task. We show that superior results can be achieved by use of kernel SVMs, as well as carefully targeted combinations of experts.

4. Method

Our approach combines shape and texture measurements in a single feature vector, with the larger weight by far given to texture. In fact, the shape measurement is only required to identify the cytoplasmatic class, and roundness of the mask (calculated as area divided by square of the perimeter) is sufficient to separate most members of this class based on a single parameter.

Texture recognition for the remaining 5 classes is far more subtle, but we note from the their descriptions that the distinctions are often ones of *scale*, rather than a specific textural pattern. This is most apparent in fine vs. coarse speckled cases, but also continues to larger spots in centromere, and even larger areas in nucleolar. We therefore seek texture measures that capture the scale at which textural variation is strongest. We also note that all the textures are completely isotropic, allowing much simpler formulations compared to the general case. To further simplify texture assessment, the images are converted from the full-colour RGB sources, dominated by the green component of the fluorescence band, to 8-bit grey scale.

Two broad types of measure contribute to the texture part of the feature vector: spectral transform and difference statistics. The frequency analysis is performed as 32-point DCT of line sections from inside the segmented mask boundaries. As the texture is isotropic, a 1-dimensional transform is sufficient to establish its frequency distribution. Transforms from all the qualifying lines within an image are averaged to reduce variability and noise, and normalised by the DC coefficient to compensate for the range of intensities within the data set. The resulting coefficients can be used for classification, with class-average profiles of the lower bins shown in Fig. 2. This illustrates some of the trends in

typical frequency distributions for each class, for example the smaller low frequency components in homogeneous and fine speckled classes, gradually rising for coarse speckled. A much sharper roll-off in nucleolar profiles relates to the large areas of bright staining combined with a small amount of fine-grain texture. Within-class spread for each bin is quite large (typically 0.3-0.5 of the average level), and so no one bin is a sufficient basis for decision, even when the average curves appear well separated.

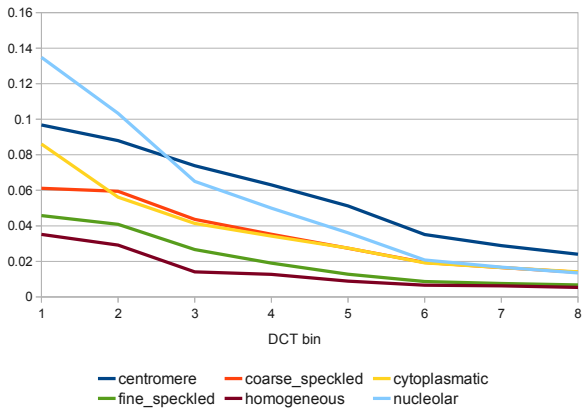


Figure 2. Average frequency profile for each class.

Pixel difference statistics are collected at different scales, to capture a broad range of textural energies. Basic average absolute difference between neighbouring pixels (horizontal and vertical combined), as defined in Eq. 1 with $\delta = 1$, is highest for fine speckled and homogeneous classes, whereas differences from 2 pixels apart ($\delta = 2$) are increased for coarse speckled and centromere.

$$D(\delta) = \sum_{i,j} |I_{i+\delta,j} - I_{i,j}| + |I_{i,j+\delta} - I_{i,j}| \quad (1)$$

Repeated subsampling by factor of 2 in each direction (following a suitable low-pass filter to avoid aliasing) and applying the pixel-difference operator creates a multi-scale textural signature. The first level of subsampling smooths out most of the finer textures, but brings the stronger gradients of centromere and nucleolar classes to pixel-level scale. Further levels of subsampling are not useful in this setting, as resulting images are too small to retain any coherent information.

The difference averages at the various scales are strongly linearly correlated with each other, but at characteristically different slopes for each class. We therefore derive the most classification benefit by taking pairwise ratios between measurements at different scales, and including them in the feature vector. The ratios are also free from dependency on overall brightness and contrast of the image - something which is difficult to achieve by explicit contrast normalisation without significant loss of discriminative information.

In summary, the feature vector consists of 1 shape measurement, 31 normalised frequency transform bins, and 8 difference statistics or their combinations. The binary intensity label supplied for each image is also included. These are passed to a multi-class SVM with RBF kernel, whose parameters are determined by cross-validation.

It is found, in line with previous works [5], that the classes which generate the greatest confusion rate are homogeneous and fine speckled. These are indeed most similar in visual appearance, with sensor noise on a homogeneous image bearing very similar characteristics to textural components of fine speckle. The problem is most acute for intermediate intensity images, where effect of noise is greater. We therefore introduce a 2nd-stage classifier expert to deal with this subset of cases. It is trained on, and applied to, only images of intermediate intensity from homogeneous or fine speckled classes, and uses a subset of features found to have the best discriminative power for this particular task. For example, it is clear that the shape parameter is not relevant to distinguishing these two classes, as are many of the coarser difference statistics. Their contribution to the feature vector is essentially noise, and their removal allows the classifier to construct a better class boundary. It is also essential to the success of this expert that only intermediate intensity images are included in the training set, as it is found that positive and intermediate cases of the same pattern class form 2 distinct clusters in feature space. Removal of interference from the positive cases allows the creation of a cleaner and simpler decision boundary for the remaining intermediate samples.

5. Results

Results reported here are for 10-fold stratified cross-validation on the contest training set, using a total of 41 features. A single multi-class SVM can achieve an error rate of 2.15%, with a standard deviation spread of 1.6% determined from 10 repeated runs. Cytoplasmatic and coarse-speckled classes are recognised perfectly (1.0 precision and recall). Remarkably, the error rate among positive samples is only around 0.5%, or typically 2 errors out of 396 positive examples. This is an order of magnitude smaller than the state-of-the-art comparable figure of 7.6% in [4], which considers sep-

aration of positive cells into 6 broadly similar classes.

So the bulk of the errors come from the darker intermediate samples, and around half of these are confusions between homogeneous and fine-speckled classes. The introduction of an expert classifier to tackle this subset reduces the overall error rate to around 1.6%.

For the purpose of discussion, error rate using shape and spectral features only (33 features) is 4.4%, and with shape and difference statistics only (10 features) it is 9.0%.

6. Discussion

The results quoted above are clearly vastly superior compared to previously published figures of above 10%. A number of factors contribute to the improved results, including a more powerful classifier type, as well as a carefully tailored 2nd-stage expert. Novel use is made of the intensity parameter as both a feature within the regular feature vector, and a selection parameter for construction of the expert training set. But the greatest contribution comes from finding measurements that characterise the specified classes and allow them to be differentiated with high accuracy. Our search was directed primarily by an understanding of the spectral content of the various textures, rejecting the far more complex state-of-the-art methods for general-purpose texture recognition. Instead, these highly targeted measurements deliver the precision needed by a clinical application at a small computational and development cost. It is worth noting that neither the spectral nor the statistical features can do this on their own, as evidenced by their relatively poor stand-alone error rates, but only through reinforcement between different parts of the feature vector.

A number of other options were considered and tested during the development of this algorithm, and it is worth recording their effects for future reference. Gamma (power law) adjustment of the input images, which can boost the brighter spots and make them more visually apparent, did not improve the discriminative strength of the feature vector. Contrast normalisation, in line with general experience, loses too much information and significantly degrades classification performance. RMS energy and variance of the difference filtered images are no more discriminative than the simple absolute mean, and higher order statistics (skewness, kurtosis) of the difference histograms also do not bring additional information. Addition of non-separable 2D spot filters does not significantly improve the detection accuracy of the centromere class. Randomised decision forests [1] were also evaluated, but could not match the accuracy of kernel SVM.

7. Further work

A clinical diagnosis cannot be made on the basis of a single cell's pattern assessment. Soda in [5] have previously considered the question of 'whole well' decision, in a voting assembly of all the cells in a slide. Unfortunately, they based their system on a very inaccurate cell-level classifier, and their attempts to improve its robustness by rejecting samples with conflicting votes or low-confidence cell decisions resulted in very high rejection rates, as well as a stubbornly high well-level error rate of around 10%. However, our more accurate cell-level classifier could form the basis of a much more reliable whole-well diagnostic system, if additional well-level labelling was available. We would also investigate the potential for combining low level features from all the cells in a sample, rather than relying solely on the class decision of cells as the only input to the whole-well diagnosis.

8. Acknowledgements

This research was supported by Engineering and Physical Sciences Research Council.

References

[1] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. Technical Report TR-2011-114, Microsoft Research, 7 J J Thomson Ave, Cambridge, CB3 0FB, UK, 2011.

[2] P. Elbischger, S. Geerts, K. Sander, G. Ziervogel-Lukas, and P. Sinah. Algorithmic framework for hep-2 fluorescence pattern classification to aid auto-immune diseases diagnosis. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on*, pages 562–565, 28 2009-july 1 2009.

[3] T.-Y. Hsieh, Y.-C. Huang, C.-W. Chung, and Y.-L. Huang. Hep-2 cell classification in indirect immunofluorescence images. In *Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on*, pages 1–4, dec. 2009.

[4] Y.-C. Huang, T.-Y. Hsieh, C.-Y. Chang, W.-T. Cheng, Y.-C. Lin, and Y.-L. Huang. Hep-2 cell images classification based on textural and statistic features using self-organizing map. In J.-S. Pan, S.-M. Chen, and N. Nguyen, editors, *Intelligent Information and Database Systems*, volume 7197 of *Lecture Notes in Computer Science*, pages 529–538. Springer Berlin / Heidelberg, 2012.

[5] P. Soda. Early experiences in the staining pattern classification of hep-2 slides. In *Computer-Based Medical Systems, 2007. CBMS '07. Twentieth IEEE International Symposium on*, pages 219–224, june 2007.